Journal of Biomedical Informatics 41 (2008) 1050–1052

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Visualizing evolution and impact of biomedical fields

Murat Cokol [a], Raul Rodriguez-Esteban [b,*]

[a] Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, Boston, MA 02115, USA
[b] Center for Computational Biology and Bioinformatics, Joint Centers for Systems Biology, Department of Electrical Engineering, Columbia University, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 22 November 2007
Available online 11 May 2008

Keywords:
Biomedical fields
Scientometrics
Scientific trends
Knowledge propagation

ABSTRACT

We describe a new tool for visualization of biomedical scientific trends. The method captures variations in scientific impact over time to allow for a comparison of relative significance and evolution of fields similar to a financial market scorecard. The tool is available at SciTrends (http://www.scitrends.net), depicting the evolution of almost 200 thousand biomedical fields in time. With millions of articles on thousands of topics published in biomedicine, we envision that only with such large-scale tools researchers can objectively understand the ever-changing interests in the biomedical sciences and make more informed decisions.

In the past five decades, biomedical research has produced millions of articles addressing various subjects in biomedicine. A significant fraction of this collective effort has been documented in the PubMed database (http://www.pubmed.com). Here, we use several informational dimensions of this database with a new method of analysis to help researchers understand the fast-paced evolution of their scientific environment. The data resulting from this analysis is available at SciTrends, a website that allows biomedical researchers to visualize the evolution and impact of biomedical trends. A practical use of such a tool is helping biomedical researchers assessing changes and opportunities in emerging fields in an age of impending 'saturated science' [1], where the percentage of resources allocated to scientific progress is expected to plateau. Our tool will help scientists get a clearer picture of the general interest in their research and evaluate their research aims and funding strategies accordingly.

A precise definition of what constitutes a scientific field is an elusive task. 'Neurosciences' is a large field in biology, however researchers that concentrate in a smaller area of 'neurosciences,' such as 'synapses,' may consider themselves in a specialty with its own structure and boundaries. Such a hierarchical organization does not allow a division of biomedical subjects into nonoverlapping fields. Moreover, interdisciplinary research that brings together different fields makes a precise parcellation difficult to make.

In an effort to help researchers find articles related to their studies, PubMed tags articles with 'MeSH terms' and 'substance names' reflecting the subjects and chemicals (or drugs) discussed in the article, respectively. The MeSH and substance ontologies capture a broad, growing scope within the biomedical sciences and are actively updated to keep track with new scientific developments. To capture the hierarchy and granularity of biomedical fields from the most general to the most specific, we adopted each MeSH term and substance name, as defined by PubMed, as a biomedical field. This way, a user can track the evolution of fields such as 'Neurosciences' or 'Synapses,' depending on the level of granularity she desires to visualize.

Our analysis covers the time period 1950–2006 and, within this time range, a total of 23,808 MeSH terms and 174,879 substance names were used for annotating the contents of 16,880,015 articles (as new data becomes available, SciTrends will be updated accordingly). We use these keywords as proxies to scientific fields, and hereafter refer to them as such. For each of these fields, we count the number of articles that mention the field each year and generate article trends.

The article trend reflects the popularity of a certain field, but gives only indirect information about the impact of the field in the scientific community. To evaluate a field's impact we utilize the journal impact factors computed by Thomson ISI (http://www.isinet.com). ISI computes impact factors only for the most influential and prominent journals [2], most journals not covered by ISI are either new or have a very low impact factor. Hence, we assumed an impact of 0 for journals not covered by ISI. This approach assigns low impact factors to journals influential yet too new to be covered by ISI but it serves as a good approximation for most journals. An alternative approach would be to exclude articles published in journals with no impact factor. This would not only reduce the coverage of our analysis but also introduce an artificial inflation of the average impact since it would introduce the assumption that articles with an impact factor are a representative sample of all articles published, while they are actually a sample skewed towards high impact. We first assign an impact value to each article as the average of its journal's impact factor

* Corresponding author. Fax: +1 212 851 5290.
E-mail address: raul@ee.columbia.edu (R. Rodriguez-Esteban).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.05.002

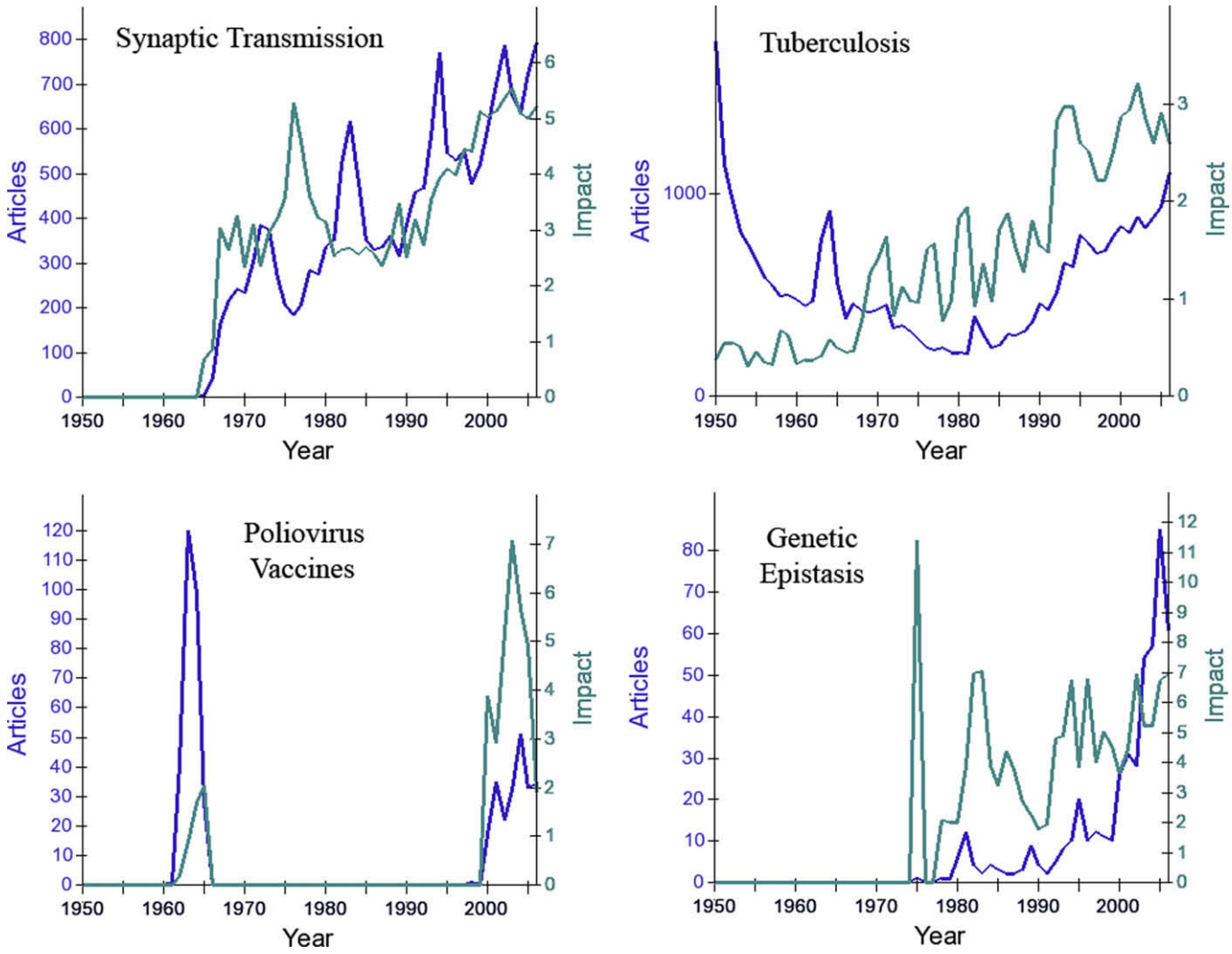

Fig. 1. Article and impact trends for 'synaptic transmission,' 'tuberculosis,' 'poliovirus vaccines,' and 'genetic epistasis.' See text for details.

values reported between 1999 and 2006. The impact values of articles are proxies for the expected citation rates for these articles, since impact factors are calculated by averaging over the citations all articles in the journal receive. We generate an impact trend of a scientific field by computing the average impact of the articles indexed with the keyword within each year.

Fig. 1 shows the article and impact trends for several biomedical fields. In the case of 'synaptic transmission,' the graphical representation allows us to see that there have been four peaks of popularity in the field and that, globally, the field is still growing. Interestingly, each of these peaks is preceded by an increase in the impact of the field, possibly due to a concomitant breakthrough. In the simpler case of 'tuberculosis,' we see that there is a regained interest of scientific community towards research in this disease, which was triggered by the appearance of resistant forms of mycobacterium tuberculosis [3]. A similar, yet more striking, trend is observed for 'poliovirus vaccines' Research in this field was nonexistent for more than twenty years after polio's 'eradication' in the sixties, but it is on the rise again [4]. In the field of 'genetic epistasis' we observe a sharp increase of interest in the last few years, reflecting the advances in high-throughput genetic interaction screens [5]. Our tool gives a bird's eye view of each field's evolution in time, which could normally be achieved only by sifting through thousands of articles.

In addition to the per-year trends, we define several global metrics of a scientific field. The total impact (I) created by a field is defined as the sum of the impact values of all the articles indexed by the keyword. We define the total number of articles mentioning the keyword as the field's volume (V). The ratio of the total impact of a keyword and its volume, the impact volume ratio (IVR), represents the expected impact generated by an article in a field. The success of the field in terms of articles and impact compared to other fields is noted in percentiles, which can be computed only by a complete knowledge of all fields in biomedicine. For 'synaptic transmission,' 17,571 articles were published until 2006, which puts it in the 98th percentile in terms of volume among all MeSH terms. IVR for this term is 3.84, which stands at 83rd percentile (Fig. 2).

We find that there is a strong positive correlation between I and V values of keywords (MeSH terms, $r^2 = 0.83$, $p < 10^{-6}$; chemicals, $r^2 = 0.90$, $p < 10^{-6}$). Using the linear correlation line, we can compute the expected impact (I_e) corresponding to the volume of a keyword. The value of I/I_e is a normalized measure of a field's success: a field with an I/I_e value larger than 1 is more successful than expected. In the case of 'synaptic transmission,' I/I_e is 1.41, meaning that articles in this field create 41% more impact than expected. We conjecture that more basic fields tend to have higher IVR and I/I_e values than more applied fields, simply because of the fact that

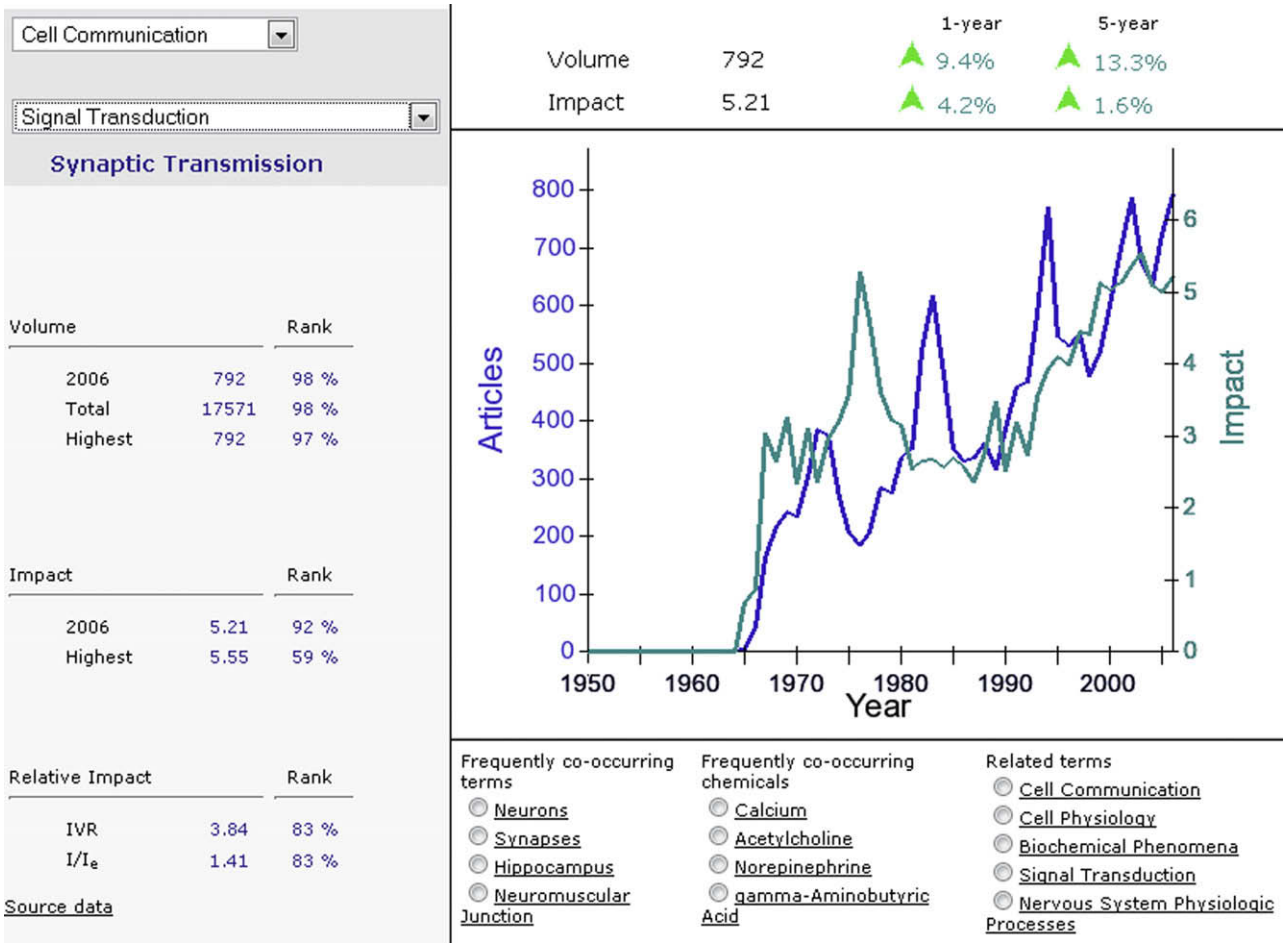

Fig. 2. SciTrends web screenshot of the data presented for 'synaptic transmission.'

applied fields tend to use (and cite) information that is generated in more basic fields of biomedicine.

SciTrends allows for single terms to be searched or browsed alphabetically. MeSH terms can be searched using a MeSH tree browser and substance terms with Chemical Abstracts Service (CAS) registry number or Enzyme Commission (EC) number can be browsed separately. The evolution of an additional term can be overlaid in a term's plot, which allows for direct comparison between trends. A list of frequently co-occurring terms and related terms is given for each term to help make a comparison analysis.

The tool described here is not predictive but descriptive. It simply allows for a visualization of trend information for specific scientific fields and it summarizes a field's popularity compared to other fields. The data represented in SciTrends provides an excellent dataset for further analysis and forecast of scientific trends, using well-established tools such as Kleinberg's burst detection algorithm [6], and are available upon request from the authors.

As previously noted for the popularity of gene names [7,8], one can see a number of different trend trajectories, including increase, stagnation, death, and even resurrection. The information dimension of impact presented in this study is analogous to stock prices, while the number of articles can be thought as market size. We believe that such an overview, using tools like SciTrends, can help science policies [9,10] and decisions of individual scientists [11] and funding organizations.

Conflict of interest

None declared.

Acknowledgments

The authors are grateful to the reviewers for their insightful suggestions and to Zeynep Gumus, Ivan Iossifov, and Chani Weinreb for comments on the earlier version of the manuscript. The authors were supported by the National Institutes of Health (training fellowship 5-T15-LM007079 to M.C. and RO1 GM61372 to Andrey Rzhetsky).

References

[1] Price DDS. Little science, big science. New York: Columbia University Press; 1963.
[2] Garfield E. How ISI selects journals for coverage: quantitative and qualitative considerations. Essays Inf Sci 1990;13:185.
[3] Iademarco MF, Castro KG. Epidemiology of tuberculosis. Semin Respir Infect 2003;18(4):225–40.
[4] Centers for Disease Control and Prevention (CDC). MMWR Morb Mortal Wkly Rep 2005;55(6):145–50.
[5] Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 2007;8(9):699–710.
[6] Kleinberg J. Bursty and hierarchical structure in streams. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining; 2002.
[7] Hoffmann R, Valencia A. Life cycles of successful genes. Trends Genet 2003;19:79–81.
[8] Pfeiffer T, Hoffmann R. Temporal patterns of genes in scientific publications. Proc Natl Acad Sci USA 2007;104(29):12052–6.
[9] Rostow WW. Why the poor get richer and the richer slow down. Austin: University of Texas Press; 1980.
[10] Tainter JA. The collapse of complex societies. UK: Cambridge University Press; 1988.
[11] Loehle C. A guide to increased creativity in research—inspiration or perspiration? Bioscience 1990;40:123–9.